Simple Linear Regression

**Regression equation**—an equation that describes the average relationship between a response (dependent) and an explanatory (independent) variable.

\[ y = mx + b \]

Deterministic Model

A model that defines an exact relationship between variables.

Example: \( y = 1.5x \)

There is no allowance for error in the prediction of \( y \) for a given \( x \).
Probabilistic Model

A model that accounts for random error.

Includes both a deterministic component and a random error component.

\[ y = 1.5x + \varepsilon \]

This model hypothesizes a probabilistic relationship between \( y \) and \( x \).

Probabilistic Model—General Form

\[ y = \text{Deterministic component} + \text{Random component} \]

where \( y \) is the “variable of interest”.

Assume that the mean value of the random error is zero \( \Rightarrow \) the mean value of \( y, E(y) \), equals the deterministic component of the model.

First-Order (Straight Line) Probabilistic Model

\[ y = \beta_0 + \beta_1x + \varepsilon \]

where \( y = \text{Dependent variable} \)

\( \beta_0 = \text{population y-intercept of the line} \)—the point at which the line intersects or cuts through the y-axis

\( \beta_1 = \text{population slope of the line} \)—the amount of increase (or decrease) in the deterministic component of \( y \) for every 1-unit increase (or decrease) in \( x \).

\( \varepsilon = \text{random error component} \)
First-Order (Straight Line) Probabilistic Model

$\beta_0$ and $\beta_1$ are population parameters. They will only be known if the population of all $(x, y)$ measurements are available.

$\beta_0$ and $\beta_1$, along with a specific value of the independent variable $x$ determine the mean value of the dependent variable $y$.

Model Development

$\beta_0$ and $\beta_1$ will generally be unknown. The process of developing a model, estimating model parameters, and using the model can be summarized in these 5-steps:

1. Hypothesize the deterministic component of the model that relates the mean, $E(y)$ to the independent variable $x$
   
   $E(y) = \beta_0 + \beta_1 x$

2. Use sample data to estimate unknown model parameters
   
   find estimates: $\hat{\beta}_0$ or $b_0$, $\hat{\beta}_1$ or $b_1$

Model Development (continued)

3. Specify the probability distribution of the random error term and estimate the SD of this distribution

   $\varepsilon \sim N(0, \sigma)$ — will revisit this later

4. Statistically evaluate the usefulness of the model

5. Use model for prediction, estimation or other purposes

Example: Reaction time versus drug percentage

<table>
<thead>
<tr>
<th>Subject</th>
<th>Amount of Drug (%)</th>
<th>Reaction Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x$</td>
<td>$y$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
Example: Reaction time versus drug percentage

**Errors of prediction** — vertical differences between the observed and the predicted values of $y$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$\hat{y} = -1 + x$</th>
<th>$(y - \hat{y})$</th>
<th>$(y - \hat{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sum of errors = 0</td>
<td>Sum of squared errors (SSE) = 2</td>
<td></td>
</tr>
</tbody>
</table>

Least Squares Line

Also called regression line, or the least squares prediction equation

Method to find this line is called the method of least squares

For our example, we have a sample of $n = 5$ pairs of $(x, y)$ values. The fitted line that we will calculate is written as $\hat{y} = b_0 + b_1x$

$\hat{y}$ is an estimator of the mean value of $y$, $E(y)$;

$b_0$ and $b_1$ are estimators of $\beta_0$ and $\beta_1$.
Least Squares Line (continued)

Define the sum of squares of the deviations of the \( y \) values about their predicted values for all \( n \) data points as:

\[
SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[ y_i - (b_0 + b_1 x_i) \right]^2
\]

We want to find \( b_0 \) and \( b_1 \) to make the SSE a minimum—termed least squares estimates

\[ \hat{y} = b_0 + b_1 x \]

is called the least squares line

Least Squares Line for Drug/Reaction Example

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( x_i^2 )</th>
<th>( x_i y_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>25</td>
<td>20</td>
</tr>
</tbody>
</table>

\[ \sum x_i = 15 \quad \sum y_i = 10 \quad \sum x_i^2 = 55 \quad \sum x_i y_i = 37 \]

\[ SS_{xx} = 37 - \frac{(15)(10)}{5} = 37 - 30 = 7 \quad SS_{xy} = 55 - \frac{(15)^2}{5} = 55 - 45 = 10 \]

\[ b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{7}{5} = 1.4 \]

\[ b_0 = \frac{\sum y_i - b_1 \sum x_i}{n} = \frac{10 - 1.4(15)}{5} = \frac{10 - 21}{5} = -0.6 \]

\[ \hat{y} = -0.6 + 1.4x \]

Formulas for the Least Squares Estimates

Slope: \( b_1 = \frac{SS_{xy}}{SS_{xx}} \) or \( b_1 = r \frac{SD_y}{SD_x} \)

\[
s_0 = SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \\ s_1 = SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}
\]

\[ s_0 = \bar{y} - b_0 \bar{x} = \frac{\sum y_i}{n} - b_0 \frac{\sum x_i}{n} \]

\[ n = \text{sample size} \]

LS Line for Drug/Reaction Example

\[ \hat{y} = -0.6 + 1.4x \]
LS Calculations for Drug/Reaction Example

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>( \hat{y} = -1 + 0.7x )</th>
<th>(y - (\hat{y}))</th>
<th>( (y - \hat{y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.6</td>
<td>1.6</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.3</td>
<td>0.7</td>
<td>0.43</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2.0</td>
<td>0.0</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2.7</td>
<td>-0.7</td>
<td>0.49</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3.4</td>
<td>0.6</td>
<td>0.36</td>
</tr>
</tbody>
</table>

The LS line has a sum of errors = 0, but SSE = 1.1 < 2.0 for visual model.

Least Squares Line—Interpretation of \( \hat{y} = -1 + 0.7x \)

Estimated intercept is negative \( \Rightarrow \) that the estimated mean reaction time is equal to -0.1 seconds when the amount of drug is 0%.

What does this mean since negative reaction times are not possible?

Model parameters should be interpreted only within the sampled range of the independent variable.

Least Squares Line—Interpretation of \( \hat{y} = -1 + 0.7x \)

The slope of 0.7 implies that for every unit increase of \(x\), the mean value of \(y\) is estimated to increase by 0.7 units.

In the context of the problem:

For every 1% increase in the amount of drug in the bloodstream, the mean reaction time is estimated to increase by 0.7 seconds over the sampled range of drug amounts from 1% to 5%.

Coefficient of Determination

A measure of the contribution of \(x\) in predicting \(y\)

Assuming that \(x\) provides no information for the prediction of \(y\), the best prediction for the value of \(y\) is \(\bar{y}\)

\[
\hat{y} = \bar{y}
\]

\[
SS_y = \sum (y - \bar{y})^2
\]
Coefficient of Determination (continued)

\[ SSE = \sum (y_i - \hat{y}_i)^2 \]

Coefficient of Determination (continued)

\[ SS_{yy} = \sum (y_i - \bar{y})^2 \quad \text{--total sample variation around mean} \]
\[ SSE = \sum (y_i - \hat{y}_i)^2 \quad \text{--unexplained sample variability after fitting} \]
\[ SS_{yy} - SSE \quad \text{--explained sample variability attributable to linear relationship} \]
\[ \frac{SS_{yy} - SSE}{SS_{yy}} = \text{explained proportion of total sample variability explained by the linear relationship} \]

In simple linear regression \( r^2 \) is computed as the square of the correlation coefficient, \( r \)

\[ r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}} \]

\( 0 \leq r^2 \leq 1 \)

Interpretation—\( r^2 = .75 \) means that the sum of squared deviations of the \( y \) values about their predicted values has been reduced by 75% by the use of \( \hat{y} \) instead of \( y \), to predict \( y \) of the least squares equation.